Analysis of April 2021 Demonstration Data for Redistricting and Voting Rights Act Use Cases

May 21, 2021
12:00 pm CT

Coordinator:
Welcome and thank you for standing by. All lines are in a listen-only mode for the duration of today's conference. Today's conference is being recorded. If you have any objections, you may disconnect at this time. I would now like to turn today's meeting over to Michael Hawes. Thank you. You may begin.

Michael Hawes:
Good afternoon, everyone, or those from points West, good morning, and welcome to the fifth webinar in our series on Understanding the 2020 Census Disclosure Avoidance System. Today's webinar will focus on an analysis of our April 2021 demonstration data specific to the Redistricting and Voting Rights Act use cases.

Our presenters today are going to be, James Whitehorne, who is the Chief of Our Redistricting and Voting Rights Data Office; Tommy Wright, who's Chief of Our Center for Statistical Research and Methodology; and Kyle Irimata, who is also in our Center for Statistical Research and Methodology.

Throughout the presentation today, if you have any questions for our speakers, go ahead and enter those into the Q&A feature of the WebEx platform. And we'll be answering those in quasi real-time during the presentations. And then we'll have some additional time at the end to answer any questions that we didn't get to. So, without further ado, I will turn things over to James to kick us off.

James Whitehorne:

Thanks, Michael, and thank you to everyone who's chosen to join us here today for this important webinar on Differential Privacy Output and the Redistricting use case. My role today is going to be, just to provide a little bit of background before Tommy and Kyle launch into the - really the meat and potatoes of today's presentation.

So, to provide that background, the Census Bureau has a statutory obligation under public law 94-171, to provide each State, the tabulation and population of that State obtained in each decennial census, and desired for the apportionment or districting of the legislative bodies or bodies of that State, and to provide those tabulations for the geographic areas for which specific tabulations of population are desired.

Census Bureau also has an obligation under Title 13 Section 9, to protect respondent data collected during the decennial census. And it's the intersection of these two obligations that has been and continues to be a priority during our drive to modernize and future-proof our disclosure avoidance techniques at the Bureau.

We're doing that through the use of differential privacy and the TopDown Algorithm. Since the start of the Bureau's move towards differential privacy, the redistricting use case has been a focus. We began by soliciting information from redistricting data users, practitioners, experts, from civil rights organizations, and from the public, for their use cases around redistricting.

Through those discussions, we confirmed as a use case the ability to draw districts with accurate population counts with race and ethnicity and voting age characteristics. Once we began developing the TopDown Algorithm, we used output from the system to start comparing its output to the published data from the 2010 census. We used congressional and State legislative districts to start.

In addition, we solicited more specific use cases, and we were able to receive 20 Section 5 redistricting plans from the Department of Justice, to add to our analysis, some of which you'll see in a moment when Tommy and Kyle go through their materials.

Although Section 5 is currently inactive, we still felt that these plans will be useful to our analysis, because they provided us examples of the smaller districts where most redistricting occurs. As we continued to improve the TopDown Algorithm, and further develop the redistricting use case, we began to address that redistricting poses a very unique challenge. It's not possible to know the configuration of the districts, until those districts are designed after the data is published.

This pushed us to seek additional ways to ensure that we're meeting the redistricting use case. Still keeping in mind the smaller districts, we began by looking at the accuracy as compared to that 2010 published data of the census block group geography.

This block group geography has some important characteristics. It's nationwide, so it provides wall-to-wall coverage for the country, and it's a small geographic area that's composed of blocks, much like small districts. We developed measures of accuracy for these block groups, and then our disclosure avoidance team combined these measures with our target of areas of 500 persons or more having a specified accuracy.

These combined measures and population threshold were translated into targets of accuracy for that TopDown Algorithm. Through our earlier discussions with redistricting data users and experts, and the Department Of Justice, we learned that most of the work of redistricting is done for jurisdictions that have a population size of 2,500 people or less.

If that's divided into four or five districts, each district will be between 625 and 500 people, respectively. Through conversations with these same groups, we also established that a jurisdiction with only 500 people is not likely to divide itself into districts.

With that in mind, we established the 500-person district size for accuracy targeting. However, since block groups benefit from a direct application to privacy-loss budget, we also include geographic areas that do not have a direct privacy-loss budget allocation, but they instead gain

their accuracy from the assemblage of blocks and other whole levels of geography that are contained within them.

These other targeted areas were places, which was both incorporated places, the legally defined places, and Census Designated Places which are not legally defined, plus Minor Civil Divisions, which are the townships and towns in those States where they're functioning governments.

It is important to note that these targets were not the only considerations as we tuned the TopDown Algorithm. We took in feedback from data users in regards to the geography, racially polarized voting analysis, previous demonstration data analysis, and effects on downstream data products accuracy.

So, with all that in mind, I'm very happy to hand this over to Tommy Wright and Kyle Irimata, who will go into more detail, what we've measured and how the current PPMS stacks up to those measurements. Thank you.

Tommy Wright:
Thank you very much, James and Michael. That's - in fact, that's a very good overview, very excellent, James. You've helped define a couple of things I was wondering about, and I'm glad you did that. The title of our presentation is "Empirical Study of Two Aspects of the Topdown Algorithm Output For Redistricting."

We focused on two things, reliability and variability. There's actually a report which is currently under review, and we hope to have it out sometime the first part of next week, or certainly by next Friday. And what we have simply done today is, we've simply looked into our draft of the report, and we've pulled out two pieces of it, and we're going to share some highlights of it.

There's a disclaimer. The views presented in this paper are those of us and not the U.S. Census Bureau. And we want to thank many people who have helped in discussions. Some are listed here, but there are many others that we need to thank.

The report has two parts, and I will just give a very high level of the technical summary. In part one of this limited study, where epsilon is 10.3, this is for the person file only. There is also housing file, which has an epsilon of 1.9 for an overall 12.2, I believe.

The question we look at in part one is, what is the minimum total ideal population of a district? And we're thinking voting districts to have reliable characteristics of various demographic groups within that voting district. So, as James mentioned, we looked at nearly 200,000 block groups across the entire country, and we're using these as proxies for voting districts.

The answer, for any block group with a total count near 600 people, the difference between what we call the TopDown Algorithm ratio of the largest demographic group, which we denote by LDG, and the corresponding swapping ratio, this is the data that were released in 2010 for the largest demographic group, is less than or equal to five percentage points, at least 95% of the time.

So this is - we want to try to clarify what is meant by this expression in part one. We also consider, in addition to - we consider places and minor civil districts as proxies for districts. A similar minimum total between 350 and 400 people, is observed for places and MCDs.

We also looked at congressional and State legislative districts across the entire country. No congressional State - no congressional or State legislative district failed our test for reliability. In part two of the study, it's essentially an update of an earlier study that we released, I believe last fall, where the previous epsilon was 4, and now we're looking at the epsilon of 10.3, which came with the April release of the public demonstration product.

Our objective here is to assess - as before, to assess the variability of the April 28th version of the TopDown Algorithm output for congressional districts and State legislative districts in Rhode Island. This is what we did before. And for three additional jurisdictions shared by the U.S. Department of Justice.

So, essentially what we're doing in part two is, we're applying the same methodology, but we're simply based on the TopDown Algorithm. Findings, given more development of the TopDown Algorithm, a larger epsilon, and additional focus on how to allocate this epsilon, we see less variability throughout, with output from the latest TopDown Algorithm. Finding second, as we reported in our earlier report, relative variability in the TopDown Algorithm increases, as we consider smaller pieces of geography and population.

Now, I will go into part one, introduction. The question again is, what is the minimum total population of a district to have reliable characteristics of various demographic groups? And all of this is very - we use data to illustrate everything we'd like to say.

For each of the 217,000 block groups in the United States, we compare closeness between the published swapping counts, based on a swapping algorithm that was used in 2010, applied to the 2010 census edited file. And we look at the corresponding TopDown Algorithm counts based on the April release, a version of the TopDown Algorithm applied to the 2010 census edited file.

Our comparisons are facilitated by a measure that we call the difference of ratios, and we represent it by DR. Definition, suppose for - we have C_{SWA} of G, and C_{TDA} of G, that these are competing counts of a demographic group G associated with the block group. And we will say more about this very shortly.

And that the total block group counts, the C_{SWA} and C_{TDA}.Then the difference of ratio is defined by this quantity. We take the swapping count for the specific demographic group, and divide it by the total count for that block group.

And then we take the count for the same demographic group based on the TopDown Algorithm and divide it by the total for that block group, based on the TopDown Algorithm. And we get the difference of these two, and we take the absolute value.

More ways of the difference of ratios, DR, for a specific demographic group G, imply that the ratios for group G due to swapping and the TopDown Algorithm in a block group, are close. Definition, when this difference of ratios, DRs and T, is sufficiently small, we say that the TopDown Algorithm count for that specific demographic group G, or ratio, provides a reliable characteristic of the block group.

So we're using the publicly available swapping data as sort of a standard of measurement. I believe that the DAS team, which is actually doing work, is working with the census edited data. So, we're using this because we wanted to use data that are available to the entire public to check and verify.

So, I want to share with you now data from a block group that happened to be a block group in Maryland, it happens to be my own block group. So, I know a little bit about my block group. From 2010, the swapping count was 1,560 people. From the latest version of the TopDown Algorithm, the population is 1,587 people.

If you look at the population 18 and over from swapping, from 2010, 1,198 people. The latest version of the TopDown Algorithm is saying 1,209 people. So if we look at some of the groups inside my block groups, some of the demographic groups inside my block group, the total Hispanic count from swapping is 133.

The corresponding count from the latest version of the TopDown Algorithm, 139. And if you compute this difference of ratios, we'll take 133, divide by 1,560, and we'll take this 139 and divide by 1,587. Get this difference, get the absolute value, and that turns out to 0.0023.

We're focusing at this format, because this is the format that we believe DOJ, the Department of Justice, has used in the past in looking at cases for redistricting for various reasons. If we look at another demographic group, the total non-Hispanic, 1,427 from the swapping algorithm, 1,448 from the latest version of the TopDown Algorithm, and the difference of ratio is 0.0023 also.

White non-Hispanic group, 1,169 from swapping, 1,185 from the latest version of the TopDown Algorithm, this measure, 0.0027. Black non-Hispanic, 36 from swapping 2010, 61 from the TopDown Algorithm. The difference of ratios 0.0154. And we complete that for the rest of the demographic groups on my block, but we can also look at the population 18 and above, and there - and so, there - okay.

We look at another block group. This is the block group in Washington, DC, from one of my colleagues. And you'll notice that the size of this block group is about double the size of my block group. And some of these difference of ratios, device, generally, you see it's sort of somewhat smaller. So, as the block group population size gets larger, the difference of ratios tends to get smaller.

So I want to look at characteristics of 12 more block groups, and these block groups span the entire country. And I want to look at these block groups in terms of their sizes. So we want to go from a very small one to a very large one. So we're going to look at these 12.

And so, we're going to look four at a time. So, in the first block group, this is a block group in Texas. It is actually the entire county of Loving County in Texas, and the population in 2010 was 82 people. The TopDown Algorithm counted 77, the latest version. The 18 and above populations in 2010, 73. The TopDown Algorithm, 75.

If we look at - and so, the specific demographic groups we look at, not all of them, but for our study, we looked at six, I believe, total Hispanic, White non-Hispanic, Black non-Hispanic, American Indian, and Alaska Native, Asian non-Hispanic, and Hawaiian and Pacific Islander non-Hispanic.

Okay. And in particular, we also look at the first - we look at the largest demographic group on a given block group. So, in this particular block group, 57 people from - according to the TopDown Algorithm, is the largest count out of 77.

The second largest demographic group happens to be Hispanic. And the third largest demographic group happens to be Asian non-Hispanic. And as I say, while we look at the top three in this display, we really only look at this in terms of our reliability characteristics, the largest demographic group on the point.

The next largest - the next block group that we look at is 500. So, the size is increasing. It is in Alabama, my home State, and the swapping count is 500. In 2010, the TopDown Algorithm said 520. If you look at the largest demographic group, again, based on the TopDown Algorithm is, White non-Hispanic. The second largest is total Hispanic. And the third largest is Black, non-Hispanic.

And you see here, the corresponding difference of ratios. The difference of ratio here is 0.215 for the largest group. The next block group is also in Alabama. It is now 1,000. And so, what you'll see here in this particular case, the TopDown Algorithm says 1,001. The 18 and above population, 745 from 2010 swapping, from the TopDown Algorithm, latest version 743.

If we look at, what is the largest demographic group for this particular block group, it happens to be Black non-Hispanic, and the count is 650. The second largest is White non-Hispanic. And the third largest is total Hispanic. The corresponding difference of ratio here for the largest demographic group is 0.0096. I apologize for the size of these, but I hope it is okay. If you get a copy of the report, that will be fine.

The fourth largest - the fourth block group that we look at now is 1,500. According to 2010 census, the TopDown Algorithm says 1,542. The largest demographic group in this particular case is Hispanic. The difference of ratios is 0.0015. The second largest is White non-Hispanic, and the third largest is Asian non-Hispanic.

Flipping, we now - the fifth block group we look at has 2,000. The largest group, White non-Hispanic. The difference in ratio, 0.0194. The next largest is 3,000. This is - and the largest

group is White non-Hispanic, 0.0131. You can see maybe a tendency for these differences in ratios to decrease as we increase the size of block group. Now, this is a 5,000, roughly 5,001, the top - the largest demographic group is White non-Hispanic.

And this - and the next group, which is in Georgia, now this is 10,000 in population. The TopDown Algorithm says 10,014. If you look at the largest demographic group is Black non-Hispanic, and the count there under the TopDown Algorithm is 4,482, and the swapping 4,475. The difference in ratios is 0.001.

And the last four, now we go to - increase further to 15,000. This is the block group again in Georgia. The largest demographic group is White Non-Hispanic, near 20,000. In Virginia, the largest demographic group, again, is White non-Hispanic.

Now we'll go to Florida. This is near 30,000. The difference in ratios is 0.003. So you sort of see this going down, and not necessarily strictly going down, but you see a tendency downward. And then the large - the very largest block group that's in the country, at least according to 2010 census, 37,452, the TopDown Algorithm says 37,303. For the largest demographic group, the difference in ratio is 0.0020.

So we can actually use these four block groups to illustrate in a very simple way, this reliable characteristic. So, if we stratify the four block groups that we just saw and maintaining the same ordering from smallest to largest into four strata, and show the difference in ratios for each stratum for the - where G is the largest demographic group, and assume that the top-down algorithm count is a reliable characteristic for the largest demographic group, if we have the following criteria.

Now, the criteria that we implement here is - for this illustration, is that the difference in ratios for the largest demographic group, is going to be less than equal to 0.005. So, trust me that in the first strata, if you go back and look, you'll see that these are the three values that we presented for the difference in ratios, and 0.0086, 0.0215, 0.0096.

And they're all - all three are larger than this criteria. So, therefore in this particular case, we would say that no block group happens to be reliable. If you look in stratum two, the values for the difference in ratios, 0.0015, 0.0194, 0.0131. And only one, which is this one, happens to be smaller than the criterion, which we have set for reliability. So one out of three is reliable.

If we look in stratum three, this is our first value. This is our second value, and this is our third value. In this particular case, in each case, they're all less than this criterion of 0.0050. So, in this particular case, we would say all three block groups happen to be reliable. Proportion points, I mean 1.00.

And in the last stratum, we have these three values, 0.007, 0.0003, and 0.002. And also in this particular case, all three happen to be reliable based on this criteria. So, back to the question. So, you're trying to answer, what is some quantity called - what is some size of a block group called? C^*_{SWA}. We use an asterisk.

And what we imagine is that we imagine taking all 200,000 block groups in the country, and we order them from smallest based on the swapping count. So this would be the largest block group population, and this would be the very largest block population.

And what we want to do is, we want to find somewhere between these two extremes in that size of a block groups, such that to the left for these smaller block groups, perhaps there's a tendency for the block groups to not provide reliable characteristics for the largest demographic group, and those block groups to the right, which would tend to be larger, we would feel comfortable saying, these block groups are reliable. So - and so, we're looking for this value, $C^*_{SWA.}$

So what did we do? Well, we took the - all of the blockers in the country, but this is looking at all - we actually don't look at all of them. We do not look at block groups that have a population less than than 50 people. And we do not look at block groups necessarily with population greater than 25 times.

So, we're looking in between. So, in this first stratum where the populations in between 50 and 99, we happen to have 128 block groups. And we're looking at three criterion here.

So the first criterion we say, out of this 128, what proportion of the block groups satisfy criteria one where the largest demographic has a difference of ratio less than equal to 0.1? It happens to be 0.1172. If we change the criterion to 0.03, that is a difference of ratio less than 0.03, out of this 128, we observed that the proportion happens to be 0.2812.

If we change the criterion to 0.05, then the proportion of these block groups that satisfy this criterion, happens to be 0.4062. Let's focus on criterion two. So, as you go - if we look at the second stratum, they happens to be 99, and you see that this proportion satisfying this criterion, happens to be 0.3030.

And as we go down, as the strata sizes increase, that is increase in terms of the size of the block group, we see that this proportion just tends to increase. There's a phenomenon going on here, which we are really curious about and want to try to explain, and there's a tendency for this to increase, not necessarily strictly increasing, but generally there's a tendency to increase.

And we say, we will stop once we get to the point where we first crossed 0.95, and that happens here. And we look across and we say, there are 8,345 block groups in this stratum, and the boundaries of this stratum happen to be between 1,050 people and 1,099 people. And so, we would - so that's what we observe.

If we look at the third criterion, in this particular case, we say, I want to continue to march down this group until I first observe a proportion of 0.95 block groups satisfying this criterion. And it happens here, and it happens in the stratum where there are 3,238 block groups. And the boundaries here are 550 people and 599. So this is where we get our 600. So, it actually - the special 0.95 happened somewhere in this interval, but we just say 600.

If you look at the rest of this table in this third criterion, you'll see that these proportions continue to increase. So, we actually have more than 95% of the block groups satisfying this criterion above that 600 threshold. So, this is totally empirical. So, using this publicly released data, we actually use in here, the one run that was publicly available.

We might say, empirically based on the data for the block groups used in our study, that for any block group with a total count near 600 people, the difference between the TopDown Algorithm ratio of the largest demographic group, and the corresponding swapping ratio for the largest demographic group, is less than or equal to five percentage points at least 95% of the time.

So it may take going through that example a couple of times, but that's what we mean by this particular statement. The TopDown Algorithm was based on one run of the algorithm. And so we wondered, what would happen if we applied the algorithm again?

So we didn't apply it just once more. We applied it actually 25 more independent times to the census edited file. And this is what we get. So we're looking for the stratum for each one, where 0.95 was first exceeded in the table.

So, imagine a table like we just went through, and we ask, from the first run, what did we see? The 0.95 was first seen, and it was seen with the proportion of 0.9589, and that happened in the stratum with the size of block groups between 550, 599.

The second run 550, 599, and the proportion was 0.9605. In every case of the independent run application of the TopDown Algorithm, we get the same interval 550 to 599, and this is our source, again, for the 600. We also - as James mentioned in his introduction, we looked at places and MCDs over 21,000.

I'm sorry, there's a zero missing here, but we looked at over 21,000 entities in this collection as an alternative to block groups. And we repeated this also 25 times with the TopDown

Algorithm, where we're looking and we're using criterion three again, and we're looking for the point at which we first call the proportion of 0.95.

In the first run, we observed that we first see 0.9621 as a proportion, and it occurs with this stratum, 329. The second run, we first see 0.95, 0.9580, and it occurs with this stratum, 250 people, the 299. And you see the results from the 25 runs.

We also looked at congressional and State and legislative districts as an alternative for block groups for the entire nation. So congressional districts, CD, State legislative districts in the upper chamber, SLDU, State legislative districts in the lower chamber SLDL. And if you want to know the counts, it's 436 congressional districts. Washington DC counts as one of these 436.

It's not - well, in terms of the State legislative districts in the upper chambers, 1,946 across the nation, and for the State legislative districts in the lower chambers 4,785. So, roughly slightly more than 7,000 entities in this category.

And what did we see? And this - with 25 different runs of the TopDown Algorithm - I want to call to your attention that the smallest entity here happens to be in the - the minimum population among these three categories of districts, happens to be with the lower chamber district, and the three.

And so, what we see when we run the algorithm is that we first get to 95. Actually, we get to 100%. In every run in this same stratum, and it's a stratum in which we have the smallest population count for a legislative district in the lower chambers across the nation. So, this is a one all the way, and that stratum is assigned as well.

Okay. So concluding remarks for part 1, remark one, C^*_{SWA} is an empirical result, and we take it to be 600. It seems to hold for block groups all across the nation for places in MCDs across the nation and for congressional and State legislative districts all across the nation.

Number two, while small demographic groups are important, in the context of redistricting, we believe, we understand that it is the largest among the demographic groups that have the potential to form districts where sufficiently large and compact minority groups have the opportunity to elect representatives of their choice.

There are many criterion, but this is certainly one that is concerned about the redistricting plans. And there are many other aspects.

Part two, and I'm going to zoom through this a little bit faster, and I hope it's not too fast, but I do want to give a flavor of just looking at data.

And so, what we will be looking at here is, we will be looking at several tables, comparing swapping counts from 2010 with the TopDown Algorithm counts using same geography. In the earlier study, and this study is actually from - using data from October 2019, the epsilon was four.

In this study, the epsilon is 10.3, and advancements have been made resulting in the April version of the TopDown Algorithm, which was released in last month data. We look at the 2010 census data for Rhode Island. We look at Rhode Island because this is where it started in the context of the 2018 planning for the decennial census.

And so, we've continued to look at Rhode Island. But we had a test site there, actually a dress rehearsal. Rhode Island has two congressional districts. It has 38 State legislative districts in its upper chamber, and it has 75 State legislative districts in its lower chamber.

We also will look at 2010 census data for three cases provided by the Department of Justice. In our early report, we actually looked at 20, but in this report, we're just limiting to three that we talked about in the main text of our earlier study.

And those, we conduct similar analyses, and we look at Panola County, Mississippi. And here,

we're trying to focus a little bit more on smaller. We look at Tate County school districts. There's 784 blocks, to give you some size. And we look at Tylertown, Mississippi, which had 136 blocks.

And I really do apologize for this. I think a little bit more time, I could have split this table up, but what we see in this first table is, looking at congressional districts in Rhode Island. So, there's two districts and the population for the first congressional district is 526,283. The second congressional district is 526,284.

The differences, by law I think, cannot - can be no more than one. And so that's held. We looked at - out of the 25 runs, we looked to see - we pulled out randomly three of them. And so, what we call Run A, the corresponding count for the first congressional district, 526,449.

For the second congressional district, 526,118. You can compare this count with this, and you compare this count with that. The second run gives similar accounts 526,173 persons, second, 526,394 in the third run. An interesting thing here.

I mean, there are many interesting things in this table, but I would just point out, for example, that if one looks at the White non-Hispanic population, 803,685, and the official apportionment from - I mean, I'm sorry, in the official redistricting from 2010, 377,109 were in congressional district one. 426,576 were in congressional district two.

You see the same pattern in the first run. 377,000, 426,000. In the second run, you see a similar pattern, 376,000, 426,000. And in the third run, 377,000 for district one, 426,000. And you can look - one can look at other things. If you look at other non-Hispanic, for example - and by the way, in this table, we not only look at counts, but we look at proportions.

So the proportion overall was 0.98 for the entire State of Rhode Island. But if we look at the congressional district, the official one from 2010, 1.61 was in this category in congressional district one. 0.34 was in the second congressional district.

If we look at the first run of the TopDown Algorithm, these proportions happen to be 1.60, and 0.35 for the second. If we look at the third, a run that we randomly selected out of the 25, 1.61, 0.35. And if we look at the third randomly selected run that we looked at, 1.61, 0.35. And one could look at other things in these tables.

The second, here we're looking at four of the districts in the upper chambers. And so, the counts are going to be smaller. So 28,161, and swapping. The corresponding count from the TopDown Algorithm latest version, 27,836. The second district, 28,079 swapping. TopDown Algorithm, 27,823. And one could look at other things as well.

In this last table, this is - let's see. This is the lower district. So, 13,000 is roughly the size of these each one. The first district, 13,881 on the swapping, 14,072 under the TopDown Algorithm. If we look at Black non-Hispanic for example, and I'm going to look at proportion, from the first district, 4.19 swapping, the proportion 4.30, TopDown Algorithm. 8.14 second district lower chamber, 8.23 for this demographic in the second district.

I want to now quickly look at - and I'm trying to watch my time, and I think I'm nearing the end. Here, we're looking at Panola County in Mississippi, and it's interesting. The population from 2010 is 34,700 people. There are many things interesting about all of these tables, but in this particular case, I think it's especially interesting because here you have White non-Hispanic proportion, population is 48.93. The Black non-Hispanic proportion population is 48.69.

The interesting part - and this is not surprising in Mississippi. I grew up in Alabama. What's interesting - well, lots of things are interesting, is that we have five districts. So, somehow these 34,700 people have to be equally divided between these five districts. And the ideal population size is 6941.4, and how to do that.

And so what you see, what was actually the official plan, is that over 50% of the population in the first district is Black non-Hispanic. Over 50% in the second district, Black non-Hispanic.

Over 50% of the White non-Hispanic in the third district. Over 50% of White non-Hispanic in the fourth district, and over 50% of Black in the fifth district.

You see the same pattern in the TopDown Algorithm results. Here, overall, this is 48.96 versus 48.93. 48.61 versus 48.69. And in terms of how this population, 34,702 versus 34,707, how does this get distributed? Well, it gets distributed in a similar way.

The Black population is over 50% in the first district, over 50% in the second population, over 50% in the third population for White non-Hispanic, for White non-Hispanic over 50% in the fourth, but also it's similar percentages, and over 50% in the fifth district for Black non-Hispanic.

So here in the official plan, 50.78. Here in this particular case, 50.82. And one could study other kinds of things in this table. I will speed up a little bit and look at, this is Tate County in Mississippi. The ideal population size is 3,764. So getting a little smaller. The TopDown Algorithm has the ideal population size 3,766.2.

And let's see. I guess - let me speed on. So the next one is a little smaller. So here, this is Tylertown in Texas. The ideal - so the population is 1,604. There are four districts. So, the ideal size of each district is going to be 402.25 people. This is from swapping.

The ideal size population in the TopDown Algorithm, this population turns out to be 1,617, and the distribution of those counts are here. And the - okay. And let's see. Do I want to say anything else? You can compare various things. You can look at Hispanic 1827. You can look at seven here. For nine here, you can look at four on the swapping, five here.

You can look at eight under the third, on the swapping. The number is I believe eight. One can see in the fourth district with higher count, eight for Hispanic 18 and above, and we see the number four. So, one can look at - so one can look throughout these tables, and they can tell you a lot. You get a feel for the data.

Okay. So, how did we actually quantify the variability? We looked at, among the 25 runs, we were concerned about, what is the variability inside the run? And I will say what that means in a minute. But we also wanted to see how the 25 runs varied relative to the swap.

So, the first - and we looked at - actually in this illustration, we actually did this for all of - many, many tables, but I'm just illustrating for the first congressional district in Rhode Island. And I'm looking at the demographic group, Asian non-Hispanic.

So the first one gave a count of 17,000. I'm sorry. Let me say, to give some context. The swapping count for this category that is in the congressional district one in Rhode Island from swapping, the count - the official count was 17,705.

But when we applied the first run of the algorithm to the data, we came away with 17,622. From the second run of the data, we saw 17,685. From the third run of the data, we saw 17,671. And you see the values that we actually observed for these.

If we average these values, we get a total of 442,120. Down here, we divide by 25, and we get roughly 17,684.8. I'm sorry if you're having some difficulty seeing this, but the paper will be better. So, what we do is we just simply look at the difference of this run minus the average of the run, square that. The difference of the second - between the second run minus the average square that difference.

So this is looking at variability among the 25 runs themselves. In this next column here, we look at values. How does this count, 17,620, compare with the swapping count? So this is - and then we do the similar thing for the second run and the swapping count, similar thing for the third run.

So at the bottom of this table here, here we come away with a total of 27,936 divided by 25. Take the square root. We get a sort of measure of a standard deviation of 33. From this column, looking at the distances, differences, the variability, we get 39.

If we compare them with swapping and we want to look at relative variation, so we take this number 33 and divide by this number, and we get this number here, 0.002. In this particular size where we're comparing variability with swapping, we take this number 39 and divide by this count, and we get a relative - sort of a relative variation of 0.002.

So this is just for one demographic group. In each table, we look at 20 different demographic groups. So we take the average of those 20 relative variations, and that's what's recorded here in this particular table. So, for Rhode Island, we have two districts. The population is that size. The average relative variations are here.

For the four districts that we looked at in the upper chamber, these are - this is the ideal population size, and this is the average of those 20 - for those 25 runs. And similarly, this is for the lower chamber. So, what you see here is that this number is going up as the population goes - this number is going up as the population of the entity is going down.

We can see this visually here. So, these first, so what's on the X axis, this is the ideal population size. So these are the four districts in Tylertown, the average variability is here. Here, we're looking at the four school districts in Tate, Mississippi, somewhat less.

Here, we're looking at Panola County, the five districts. And you can see this period. Here, this table, we look at, are things - what is the variable - what's the difference within the variability before and the variability now? So, these Xs represent variability before, and these circles represent the corresponding variability now. And so, in each case, you actually see that the variability has decreased. I think we have a little time. Thank you.

Michael Hawes:
All right. Thank you, Tommy. And if I could go back one here. All right. So before we take some additional questions, I do want to say, if anybody wants to learn more information about

our disclosure avoidance modernization efforts, and our development of the TopDown Algorithm, please check out our Web site.

You can just go to census.gov and search disclosure avoidance. We have a lot of great resources there, frequently asked questions, issue papers, videos, and a whole lot more. So, definitely check that out. And if you want to stay updated on our latest developments, definitely subscribe to our newsletter. You can do that from our disclosure avoidance page at census.gov. We send out messages every week or two with all the latest developments and news.

And so, with that, we have about eight minutes left to answer some additional questions. So I'm going to introduce Meghan Maury, who is a senior Advisor in Our Office of the Director, and she's going to help moderate some of these questions that we didn't get to in the chat. So Meghan, take it away.

Meghan Maury:
Thank you so much, Michael, and so excited to be here. There were very many simple questions that I think we can just go through really quickly. Simple questions like, will the slides be available? The answer is yes. They'll be available on our event Web site after the event, and we hope that you go take a look.

Second question is - that we're hearing pretty commonly is, can we see how the TopDown Algorithm performed in comparison to the unswapped data? And Michael, you did answer that in the Q&A, but if you wouldn't mind just telling folks here so - if anyone didn't see it.

Michael Hawes:
Yes, absolutely. And that's a very important distinction to note. So, the analysis that Tommy was presenting there, they were showing deviations from the published 2010 data. Some of the errors that’s observable in those will be due to the swapping routines that were used in 2010.

For those of you who aren't aware, for the 2020 censuses, we infused noise into the published

data using a mechanism known as swapping in order to protect privacy. And unfortunately, with traditional methods like the swapping algorithm used for those censuses, you need to keep the parameters of those methods and their impact on data accuracy confidential, in order to preserve the integrity of those privacy protection routines.

So, while the tuning that we've been doing for the Disclosure Avoidance System has been using the unswapped data as our comparison, we are only able to publicly share analysis on accuracy and fitness for use in - as comparisons to the swapped data.

So always bear in mind that some of the differences that you see between the published 2010 data and the demonstration data, are going to be due to the error introduced with - from that 2010 swapping mechanism.

Tommy Wright:
Thank you for that, Michael. And if I could add, we - I think we discussed this quite a bit, in fact, over several years, what should we be looking at, the CEF - as enumerated population or not? I think we wanted the public to have an opportunity to try to reproduce what we've done.

And so, one way to facilitate that is every - most of everything, except for the 25 runs, the 25 runs are not publicly available, but the - most of what's in our presentation is available to the public so that people can easily reproduce what we've done.

Meghan Maury:
Yes. That's helpful context too. I really appreciate that. There were a number of questions that are all along the same lines that said, you know, why did we use block groups? Why not use VTDs, which I think stands for voting tabulation districts, if I'm right. Why did we show the particular block groups we did as examples?

James, I know you had spoken to the VTD question. I think the answer to that really important

to, again, make sure that everyone hears. But if folks want to address the other sort of geography-based questions, that would be great too.

James Whitehorne:
Yes. I mean, did you want me to reiterate my response about the VTDs? The VTDs that we have in our census universe from the 2010 census, so the ones that we could use in a project like this, where we're working with the 2010 data, are a great national representative set of data.

We have lots of idiosyncrasies in that data. There was an experiment that was conducted in California, where they merged multiple districts, fire districts, library districts, water districts, voting districts, and then they put those pieces into the dataset.

So, analyzing those would not really provide actionable data, because it's not actually representing voting districts themselves. We have several States that have what we call pseudo districts, which is where t- in their requirements for how they provide their updates to the Census Bureau, they have to use existing census features, but their voting districts don't follow existing census features.

So they give us what we call the pseudo district. So it's not actually - it's an approximation of the district. It's not an actual district. And so, we have some of those. And then in other cases, those pseudo geographies or pseudo VCDs, are built where they create sort of super precincts.

And so, these are the overly large and overly representative of what the individual precincts are. So, it's just not a nice even national set of data from which to do an analysis like this, so that you can make really concrete conclusions from doing that study.

Tommy Wright:
And we actually did start, in an attempt to try to use those data, and then decided not to, for the various reasons that James has mentioned.

Meghan Maury:
Yes, that's really, really helpful. And Tommy, am I right, that you ended up using block groups because of, again, that national coverage?

Tommy Wright:
National coverage.

Meghan Maury:
They approximate the same - go ahead, please.

Tommy Wright:
The national coverage easy access and understandable by people as to what we were talking about, but for various reasons, and this was - as I say, this wasn't a single choice. It was just discussed quite a bit, but we settled on block group, but you're right. It has a nice over 200,000 entities that people can identify with and the sizes - we wanted to focus on smaller types of geography, and it just seemed to have been convenient to use the block groups.

Meghan Maury:
That makes a ton of sense. There were other somewhat technical questions that asked, you know, was - is the data you used in this, the same as what was published in the PPMF? And Michael answered that in the Q&A that for the most part, this is - and Tommy just answered it as well, but many pieces of this are based on what's available publicly, but they did do those sort of 25 runs of the TDA to examine that variability.

And I think that that leads to one other question we didn't quite answer as much in the chat, which is, what is the difference between one run of the TDA and the next?

Tommy Wright:
So, the noise - in my understanding, and I don't understand much, so maybe I should defer to Michael, but my understanding is that there is a probability distribution, which generates noise,

which is added to accounts. And so, one run means that we make use of this probability distribution to generate noise in various ways to various counts nationwide at a very high level.

And so, we just repeat that 25 different times. So a run in my mind, means that we're adding noise throughout the entire country, with lots and lots and lots of generations from this probability distribution. And I yield to Michael.

Michael Hawes:
Yes, I'll add on to what Tommy was saying. So, put most simply, a run in this context is taking the exact same algorithm, with all of the exact same settings, the same privacy-loss budget, holding it exactly the same and just operating it multiple times.

And so, the difference that you get there is going to be the difference in the random numbers that drive the noise infusion itself. So Tommy mentioned the probability distribution. So when the noise is added, we pull a random number.

We have a cryptographically secure random number generator. And you generate a random number for every query that you're performing against the confidential data. And using that random number, you then select from a probability distribution centered around zero, you select the amount of statistical noise or uncertainty that's going to be injected or added or subtracted from each and every one of those statistics you're calculating.

Now, the most likely circumstance would be you pull a zero and you inject no noise. With slightly less likelihood, you'll pull a one or a negative one. With even less likelihood, a two or a negative two and so on. And the shape of that probability distribution is determined by your privacy-loss budget.

But underneath all of that are the random numbers that determine where in that probability distribution you're actually pulling the amount of noise to add or subtract. And so, even with the

exact same algorithm, the exact same settings, you're going to get variation from one run to the next, just by virtue of the difference in the random numbers that get generated.

So the idea of using 25 different runs of the algorithm, is to assess not just the variability due to the settings of the system, you can do that with one run, but by using the 25 runs, you're able to assess the variability of one run to the next, because the luck of the draw, you might get different random numbers that might yield a different result. And so, it's really helpful to see what that inherent variability due to the random numbers might be.

Meghan Maury:
Okay. One last question before we jump off. Many people in the Q&A asked different questions about other analyses of the PPMF. And I will say that there are many analyses that people internally looked at to - as they were doing this work on the algorithm.

There's also a number of analyses externally that people have done on the PPMF to make judgments about things like racially polarized voting analysis, and that sort of thing. Many of the analyses have been provided to us, and of course, we read all of them. So, we're taking a look at that as part of our feedback structure.

And of course, when our analyses, when - if our analyses are published for folks, we will make sure to call your attention to them, like Tommy's fantastic research here. Anything else you want to wrap us up with, Michael?

Michael Hawes:
Yes. And I just want to add to what you were saying Meghan. Yes. So, all of the feedback, all of the analyses that have been provided to us so far, and that we'll continue to get through the end of our feedback period on the 28th, all of that is being reviewed by a wide team of us here at the Census Bureau.

And we'll be kind of distilling that and reflecting on all of that input to inform the upcoming

decision-making in June about the final parameters and final privacy-loss budget. They will be used to produce the redistricting data products.

And it's critical to note there that no final decisions have been made about that privacy-loss budget or those parameters yet. The April demonstration data was merely the results of the tuning experiments that we did. But we've already gotten actionable feedback from many of our data users that have yielded additional changes that we're intending to make, that we're currently investigating right now, in terms of that allocation of privacy-loss budget from - for one geographic level or another, or for one set of tabulations versus another.

And so, all of this feedback that we're getting, these analyses that external experts are performing and submitting to us, those will also further inform additional changes, additional kind of tweaking of those privacy-loss budget allocations, tweaking the overall setting of the privacy-loss budget, and much, much more.

So please, do your own analysis. Submit that feedback to us, because it will be enormously helpful for our decision-makers when they have to make the difficult decisions about, what's the appropriate level of privacy protection? What's the appropriate level of accuracy, and how do we ensure fitness for use for the priority use cases?

So, with that we are over time. I want to thank everybody who participated today, everybody who attended. You had great questions coming in, great comments coming in via the Q&A. Again, if you want to stay informed with what we're doing, check out our Web site, subscribe to our newsletter.

We'll have more information coming over the next weeks, and we look forward to receiving any additional feedback that any of you have on your own analyses. So with that, I will turn things back over to the operator, and thank you all very, very much.

Tommy Wright:
Thank you.

Coordinator:
Thank you. That concludes today's conference call. Thank you for your participation. You may disconnect at this time.

END